storage medium for storing an information categorizing process software program.

Page 1, line 33 to page 2, line 2:

Finding a set of pieces of information having a property of similarity from a vast amount of information is called a "clustering". The clustering, which is a well-known technique in the information processing field, is widely used to categorize a large amount of documents.

Page 2, line 14 to page 4, line 9:

FIG. 1 is a block diagram of a first embodiment of the present invention, showing the construction of an information categorizing apparatus that performs a clustering process on a search result provided by one search service.

FIG. 2 shows a plurality of documents as a result of search results provided by a search service used in the first embodiment.

FIG. 3 is a block diagram showing the construction of a clustering processing unit shown in FIG. 1.

FIG. 4 is a flow diagram diagrammatically showing the steps of a document categorizing process in the first embodiment.

FIG. 5 shows the content of a feature table illustrating the relationship between a feature extracted from the title of each document and the document having the title containing the feature.

FIG. 6 shows a categorize result of each document based on the feature table shown in FIG. 5.

FIG. 7 shows a clustering result of document titles based on the categorize result shown in FIG. 6.

FIG. 8 is a block diagram showing the construction of an information categorizing apparatus that clusters the search result provided by a single selected search service.

FIG. 9 is a block diagram showing the construction of an information categorizing apparatus that clusters the search results provided by a plurality of search services.

FIG. 10 is a block diagram showing a second embodiment of the present invention.

FIG. 11 shows clustering results that have been obtained by clustering a plurality of documents resulting from the search by a search service.

FIG. 12 is a flow diagram diagrammatically showing information categorizing process steps in accordance with the second embodiment of the present invention.

FIG. 13 shows results that have been obtained by subjecting the clustering result shown FIG. 11 to a cluster order rearranging process.

FIG. 14 shows the construction of a third embodiment of the present invention.

FIG. 15 is a flow diagram diagrammatically showing information categorizing process steps in accordance with the third embodiment of the present invention.

FIG. 16 shows a clustering result shown in FIG. 11 and a summary table thereof.

FIG. 17 shows a clustering result that has been obtained by clustering URL addresses and a summary table thereof.

Disclosure of the Invention

To achieve the above object, an information categorizing method of the present invention includes a step of acquiring a plurality of search results searched by a search service through a clustering module, a step of performing a clustering process on the search results through the clustering module, and a step of outputting the clustering result from the clustering module.

The information categorizing method may further include a step of converting, through a converter module, the search result searched by the search service into a format that is processed by the clustering module.

The converter module is arranged correspondingly to each of a plurality of search services when the clustering process is performed correspondingly to the plurality of search services.

A search process may be performed using one search service selected from the plurality of search services and the clustering process may be performed on the search result searched by the selected search service. Search processes may be performed in parallel using at least two search services of the plurality of search services, respective search results may be collected, and the clustering process may be performed on the collected search results. Search processes may be performed in parallel using at least two search services of the plurality of search services, and the clustering process may be individually performed on the search results.

When the clustering process is performed on the search result, information to be clustered is at least one of the title of a document, a URL address, an update date, and a file size of an individual search result.

In the information categorizing method, the order of cluster of a clustering result may be rearranged using a score indicating the degree of match between the clustering result and a search request for each document and the clustering result with the cluster order thereof rearranged may be then output.

Page 5, line 29 to page 7, line 4:

An information categorizing apparatus of the present invention includes a clustering module for acquiring a plurality of search results searched by a search service, performing a clustering process on the search results, and outputting the clustering result.

The information categorizing apparatus may further include a converter module for converting the search result searched by the search service into a format that is processed by the clustering module.

The information categorizing apparatus may include a cluster order setting module which rearranges the order of cluster of a clustering result using a score indicating the degree of match between the clustering result and a search request for each document and outputs the clustering result with the cluster order thereof rearranged.

The information categorizing apparatus may further include a summary table generator unit for generating a clustering result summary table indicating the summary of the clustering results based on the clustering result, and a display control unit for outputting the clustering result summary table together with the clustering result.

A storage medium of the present invention stores an information categorizing software program in which a clustering module performs a clustering process on a plurality of search results that have been searched by a search service in response to a search request of a user, and outputs the clustering result. The information categorizing software program includes a step of acquiring the search result from the search service, a step of performing the clustering process on the acquired search result and a step of outputting the clustering result.

The step of performing the clustering process may be performed subsequent to a step of converting the search result searched by the search service into a format that is processed by the clustering module.

The information categorizing software program may include a step of rearranging the order of cluster of the clustering result using a score indicating the degree of match between the clustering result and a search request for each document and a step of outputting the clustering result with the cluster order thereof rearranged.

The information categorizing software program may include a step of generating a clustering result summary table indicating the summary of the clustering results based on the clustering result, and a step of outputting the clustering result summary table together with the clustering result.

Best Mode for Carrying out the Invention

The embodiments of the present invention are now discussed. The discussion of the embodiments that follows not only covers the information categorizing method and the information categorizing apparatus but also the specific process content of the information categorizing process software program of the present invention stored in the storage medium.

(First embodiment)

FIG. 1 shows a first embodiment of the present invention, including, as major components thereof, a search service 1, a converter module 2, and a clustering module 3. The converter module 2 and the clustering module 3 in combination corresponds to an information categorizing apparatus.

Page 8, line 12 to page 9, line 5:

Referring to FIG. 3, the clustering processing unit 33 includes a feature extractor 331, a feature table generator 332, a document categorizer 333, a document categorize result memory 334, an output controller 335, and a display unit 336. The feature extractor 331 extracts features from the result of the morphological analysis provided by the morphological analysis unit 32.

The feature table generator 332 generates a feature table indicating the relationship between the features extracted by the feature extractor 331 and the documents D1-D7. The feature table will specifically be discussed later.

The document categorizer 333 references the above-mentioned feature table, thereby grouping the documents D1-D7 into a plurality of clusters from the standpoint of semantical similarity. Specifically, based on the features contained in the titles T1, T2,..., T7 of the respective documents D1, D2,..., D7, documents having the same feature in common are treated as one group, thereby forming one cluster. The document categorizer 333 may contain a synonymous feature dictionary (not shown). To group the documents having the same feature in common into a cluster, the document categorizer 333 may determine a

common feature referencing the synonymous feature dictionary for the presence of any synonym. When there is a synonym, the document categorizer 333 may include the corresponding document into the same cluster.

The document categorize result memory 334 stores the content categorized by the document categorizer 333. The output controller 335 reads the content of the document categorize result memory 334, and displays the content on the display unit 336.

B 5

The information categorizing process steps of the present invention performed in the above-referenced arrangement are now discussed. The information categorizing process steps of the present invention are roughly shown in a flow diagram in FIG. 4. Specifically, a step of acquiring a search result provided by a general-purpose search engine is performed (step S1), a step of performing a clustering process on the acquired search result is performed (step S2), and a step of outputting the clustering result (step S3) is performed. The information categorizing process steps are now discussed in more detail, referring to specific examples.

Page 9, line 35 to page 10, line 32:

The document categorizer 333 references such a feature table to cluster the features. The categorizing result is shown in FIG. 6.

The categorizing result is also stored in the document categorize result memory 334. In the document categorize result shown in FIG. 6, reference is made to a cluster (containing documents D1, D4, D6, and D7) categorized according to the "sheet". As shown in FIG. 2, the document D1 relates to the sheet cassette, the document D4 relates to the sheet setting, the document D6 covers the smearing of sheets through printing, and the document D7 relates to the mounting of the sheet cassette.

B 6

In this way, all documents D1, D4, D6, and D7 relate to the sheet. There will be no problem if these documents are grouped in the same cluster, and the categorize result is deemed appropriate.

As for the clusters categorized according to the feature "cassette" (including documents D1, D4, and D7), the document D1 relates to the sheet cassette, the document D4 relates to the sheet setting, and the document D7 relates to the mounting of the sheet cassette, as described in the documents shown in FIG. 2.

All documents D1, D4, D6, and D7 cover the setting of sheets. There will be no problem if these documents are grouped in the same cluster, and the categorize result is deemed appropriate.

Reference is made to the cluster (containing documents D2, D3, D5, and D7) categorized according to the feature "mounting". As shown in FIG. 2, the document D2 relates to the mounting of an expansion memory, the document D3 relates to the mounting of an interface card, the document D5 relates to the mounting of a hard disk, and the document D7 relates to the mounting of a sheet cassette.

B6

All documents D2, D3, D5, and D7 relate to the mounting of something. There will be no problem if these documents are grouped in the same cluster, and the categorize result is deemed appropriate.

The reason why such an appropriate categorizing is performed is that the features are extracted from the document titles, and that the documents are categorized according to the features. The writer of each document typically conveys the main point of the document in the title of the document. Categorizing the documents using the features contained in the title of each document prevents the categorize result from becoming discursive and lowers the possibility of generating a noise cluster. Since the writer of each document conveys the main point of the document in the document title, the categorizing focuses on the viewpoint of the writer of the document.

Page 11, lines 26–29:

B7

FIG. 8 shows the construction of an information categorizing apparatus for performing the above-referenced clustering process using a plurality of search services. There are available three search services of a first search service 1a, a second search service 1b, and a third search service 1c.

Page 13, line 33 to page 14, line 30:

The process program for performing the information categorizing process in this embodiment may be stored in storage media such as a floppy disk, an optical disk, and a hard disk. Such storage media fall within the scope of the present invention. The process program may be acquired through a network.

B8

(Second embodiment)

A second embodiment of the present invention is now discussed.

As discussed in connection with the first embodiment, the clustering method of extracting a feature from the title of a document is excellent in terms of the amount of computation and process time and permits appropriate clustering. Since the amount of information to be clustered is relatively small for the overall volume of each document, the entire document is not always properly clustered. A title may not properly represent the content of the document, or an inharmonious title largely unrelated to the content of a

document may be used. In such a case, clustering accuracy is substantially degraded with no good clustering result expected.

The clustering method based on the extracted feature checks the frequency of occurrence of the feature, and then automatically categorizes the documents for clustering. Since such a clustering process does not parse the document, the resulting clusters (a set of documents derived through the clustering process) are not necessarily a set of documents having semantic similarity.

B 8

Even in such a case, information categorizing preferably presents a clustering result, satisfying the search requirement of the user.

In this embodiment, the search result obtained from a general-purpose search service is subjected to the clustering process, and the cluster order of the clusters derived through the clustering process is rearranged. The clustering result is thus presented to the user in a manner that meets the search requirement of the user.

The second embodiment of the present invention is now discussed.

FIG. 10 shows the construction of the second embodiment of the information categorizing apparatus of the present invention. Referring to FIG. 10, there are shown, as major components, a search service 101, a converter module 102, a clustering module 103, and a cluster order rearranging module 104. The converter module 102, the clustering module 103, and the cluster order rearranging module 104 in combination corresponds to the information categorizing apparatus. In particular, this embodiment is characterized by the cluster order rearranging module 104.

Page 16, lines 9-15:

B 9

FIG. 12 diagrammatically shows a flow diagram of the information categorizing process steps of this embodiment. A search result searched by the search service 101 is acquired (step 12S1), the clustering process is performed on the acquired search result (step 12S2), and the clustering result is output (step 12S3). The cluster order of the clustering result is rearranged (step 12S4), and the rearranged clustering result is output (step 12S5). The information categorizing process is discussed in more detail, referring to a specific example.

Page 20, lines 13-35:

B 10

(Third embodiment)

A third embodiment of the information categorizing apparatus of the present invention is now discussed.

When the number of clusters obtained through clustering is not so large in the information categorizing process, learning all clustering results does not take much user's time.

The number of clusters obtained through the clustering process becomes occasionally large up to several tens to several hundreds. In such a case, even merely viewing all clustering results requires a great deal of attention.

B 10

In the third embodiment of the present invention, the clustering process is performed on the search result provided by a general-purpose search service, and a table for allowing the user to glance at the summary of the clustering results obtained through the clustering process is formed. In this way, the user can efficiently search for his desired information.

The third embodiment is now discussed in detail.

FIG. 14 diagrammatically shows the third embodiment of the present invention. Referring to FIG. 14, there are shown a search service 141, a converter module 142, a clustering module 143, a clustering result summary table generator module (hereinafter referred to as a summary table generator module) 144, and a display control module 145. The converter module 142, the clustering module 143, the summary table generator module 144, and the display control module 145 in combination correspond to the information categorizing apparatus. In particular, the third embodiment is characterized by the summary table generator module 144.

Page 22, lines 20-35:

B 11

The information categorizing process in the third embodiment thus constructed of the present invention is discussed. FIG. 15 diagrammatically shows a flow diagram of the information categorizing process steps of this embodiment. A search result searched by the search service 1 is acquired (step 15S1), the clustering process is performed on the acquired search result (step 15S2), and the clustering result is output (step 15S3). A step of generating a summary table is performed based on the clustering result (step 15S4), and the generated summary table is displayed together with the above-mentioned clustering result (step 15S5). To display the generated summary table together with the above-mentioned clustering result, the summary table may be superimposed on the clustering result on a screen. Alternatively, the summary table and the clustering result are separately arranged so that the display unit displays the summary table followed by the clustering result. When the clustering result is large in volume, the user may scroll through the clustering result to successively see it.